Multivariate visualization by scatterplots

Multivariate visualization using scatterplots

Fabrizio Lamberti[1], Federico Manuri[2] and Andrea Sanna[3]

Politecnico di Torino, Italy

[1]fabrizio.lamberti@polito.it

[2]federico.manuri@polito.it

[3]andrea.sanna@polito.it

Synonyms

Scatter chart; scatter diagram; scatter graph; scatterchart; scattergram; scattergraph; scatterplot

Definition

Multivariate visualization by scatterplots is the usage of diagrams to visualize sets of data that have more than three variables.

A scatterplot is a chart or mathematical diagram displaying a set of data as a collection of points using Cartesian coordinates, usually defined by horizontal and vertical axes. Each point on the chart represents two variables, x and y, calculated independently to form bivariate pairs (x_i, y_i). A functional relation between x and y is not necessary. The purpose of a scatterplot is to reveal (if existing) the relation between the displayed variables.

Introduction

Multivariate visualizations deal with the challenge of displaying sets of data with three or more variables: this peculiar feature poses two kinds of problems. First, most of the charts and diagrams usually adopted to visualize data cannot display more than three dimensions adequately. Second, the effectiveness of the visual effects adopted to represent different variables deteriorates when the number of variables increases.

Scatterplots may be considered, among the different types of data visual representations, as one of the most useful and versatile, especially in statistics. According to (Miller, 1995), the term first appeared as Scatter Diagram in a 1906 article in Biometrika, “On the Relation Between the Symmetry of the Egg and the Symmetry of the Embryo in the

Springer

Frog (Rana Temporaria)" by J. W. Jenkinson. However, the term only came into wide use in the 1920s when it began to appear in textbooks, e.g. F. C. Mills, Statistical Methods of 1925. The Oxford English Dictionary gives the following quotation from Mills: "The equation to a straight line, fitted by the method of least squares to the points on the scatter diagram, will express mathematically the average relationship between these two variables". Figure 1 provides an example of scatterplot diagram.

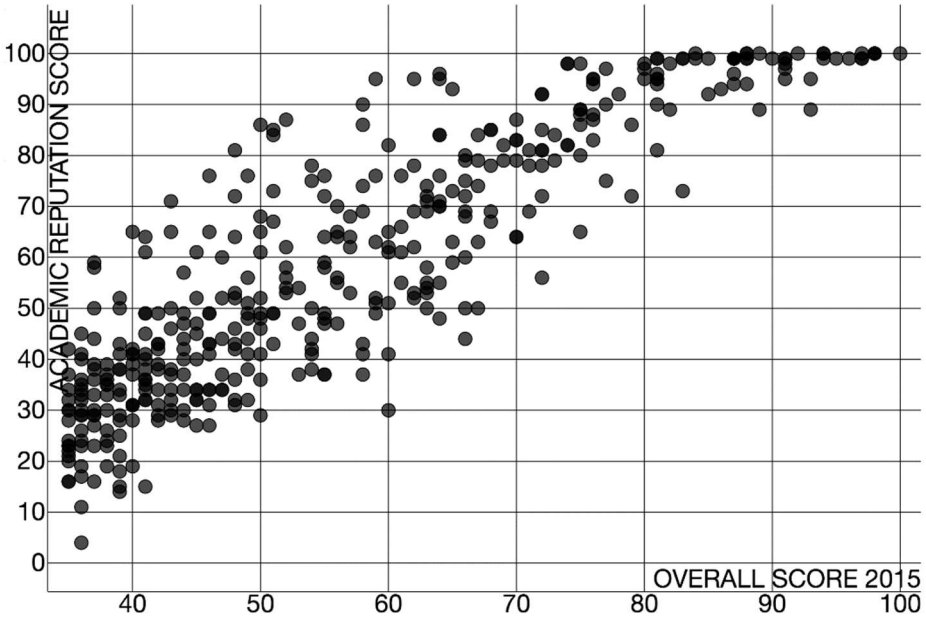

Figure 1: An example of scatterplot diagram.

Scatterplots are mainly appreciated for their ability to reveal nonlinear relationships between variables. Moreover, scatterplots are typically used to identify correlations between variables, with a certain confidence interval. Another usage for the scatterplot is to compare similar data sets. Since the main problem of multivariate data is to correctly understand and analyze them, pointing out relationships, patterns or outliers, a scatterplot provides a suitable visualization tool for multivariate data due to its intrinsic features.

Usage

Different scenarios lead to different tasks when dealing with multidimensional visualization techniques. As defined by (Valiati, 1995) and further described by (Pillat et al., 2005), five major tasks can be considered as objectives a user might want to fulfill when using a visualization tool to display or analyze multivariate data: identify, determine, compare, infer and locate. Scatterplots can be used to assess all these different tasks and have been applied to data in many different fields of use, such as automotive, finance, pharmacology, environment, weather forecast, telecommunication, food and many others.

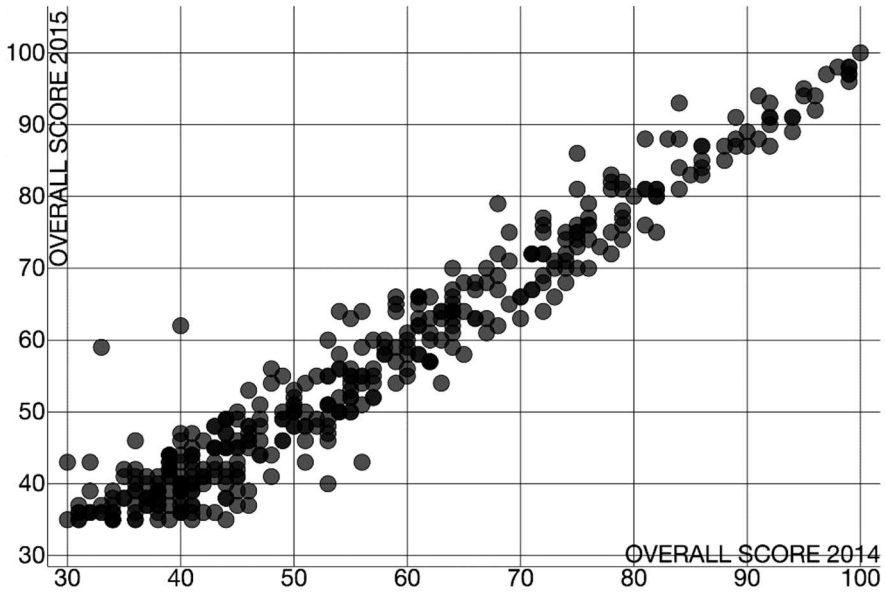

Figure 2: This scatterplot suggests a correlation between the two displayed variables.

Identify

This task refers to any action of finding, discovering or estimating visually:

- properties like symmetrical or asymmetrical distribution, values or dispersion;
- correlation, data dependency or independency;
- similarities or differences;
- clusters as a result of similarity, continuity, proximity or closed shapes;
- thresholds, patterns, data variation.

The Identify task takes place anytime the user analyzes the chart with the purpose of finding, estimating or discovering new information about the data. The task ends when the user finds the information he/she was looking for or the current goal changes. Figure 2 shows an example of scatterplot that clearly suggests a linear correlation between the displayed variables.

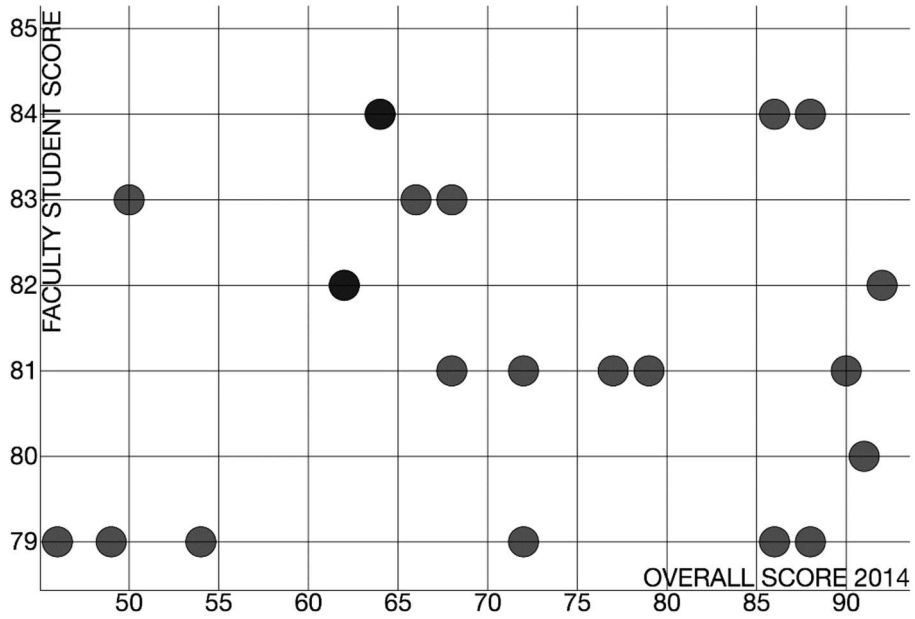

Figure 3: a scatterplot visualization that simplify the computation of the mean.

Determine

This task corresponds to the action of calculating, defining or precisely designating values such as:
- mean, median, variance, standard deviation, amplitude, percentile;
- sum, differences, proportions;
- correlation coefficients, probabilities or other statistics such as hypotheses test.

This task begins when the user needs to calculate a specific value and ends up when the calculation is completed. Figure 3 shows a scatterplot that allows to derive the precise value of each point in order to compute precise calculations such as the mean value.

Compare

This task takes place when the user wants to compare data that have been previously identified, located, visualized or determined. The user may compare data to analyze dimensions, data items, clusters, properties, proportion, values, locations and distances or visual characteristics. The Compare task is an analytic task the user performs specifically if he/she compares data items displayed in the graphical visualization. Figure 4 shows a scatterplot configuration that enhances the comparison task.

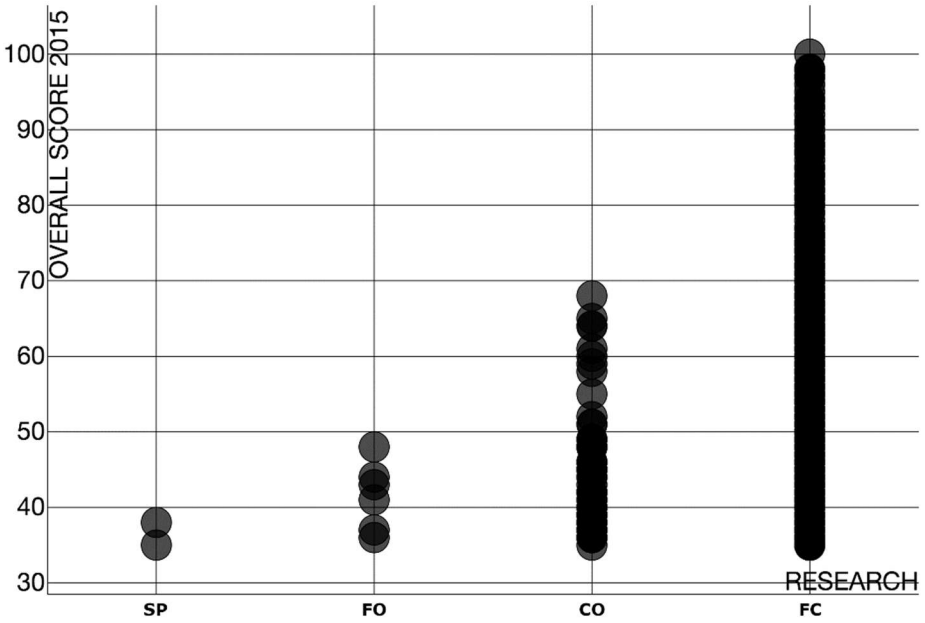

Figure 4: a scatterplot configuration that enhances comparison.

Infer

This task refers to the action of inferring knowledge from the visualized information, such as defining hypotheses, rules, probabilities or trends, attributes of cause and effect. This task usually takes place after determining, identifying or comparing information and it is performed as part of the mechanism of data analysis, thus it may not be completed at once, requiring consecutive applications of the other visualization tasks. Analyzing Figure 1 it is possible to infer a hypothesis, e.g. that the y variable is the cause of the trend of the data.

Locate

This task refers to the actions of searching and finding information in the graphic representation: they can be data points, values, distances, clusters, properties or other visual characteristics. The task begins when the user starts examining the visual representation and finishes when he/she recognizes the desired information. Figure 5 shows a scatterplot visualization that enhances the identification of outliers.

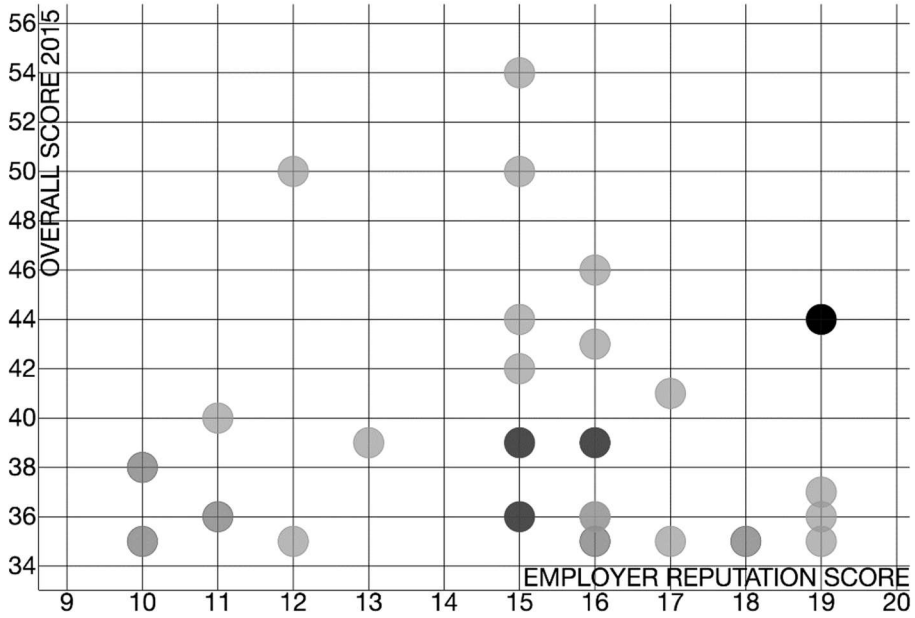

Figure 5: a scatterplot visualization that enhances the identification of outliers.

Dimensions

The main problem when using the scatterplot to visualize multivariate data is that its basic version is limited to only two variables, thus making it difficult to correctly visualize and analyze all the data. In order to overcome this problem, different solutions have been proposed through the years to enhance the scatterplot.

Adding dimensions

Even if the basic scatterplot may display only two variables, various techniques have been researched and adopted through the decades to increase the dimensionality of scatterplots by one, two, or even several additional dimensions. A bidimensional planar

scatterplot of two variables X and Y can display additional variables by correlating them to one or more graphical features of the plotted points.

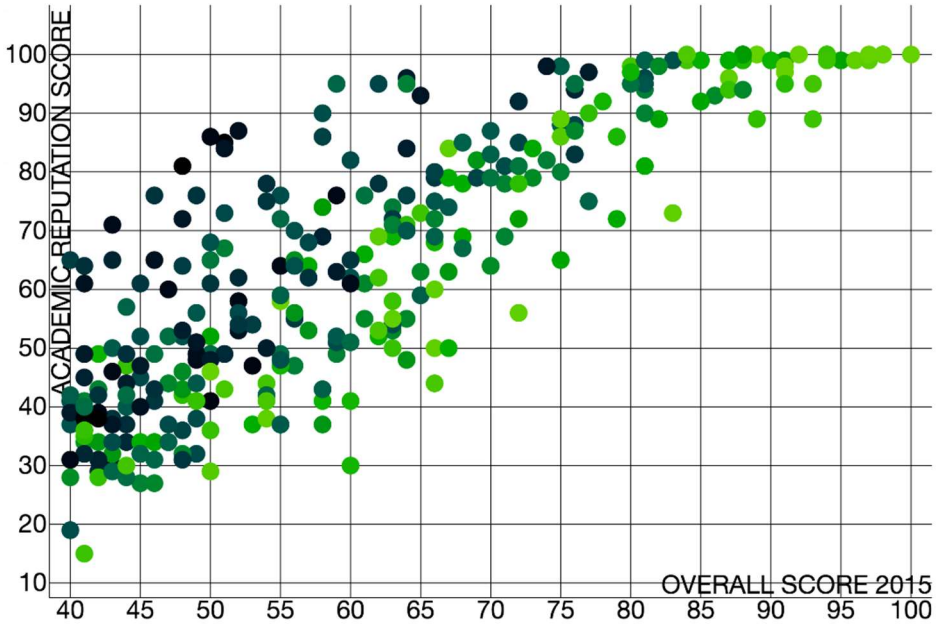

Figure 6: a scatterplot with an additional variable visualized as color.

Color

One approach is to show a third dimension through a color map. Colored points on a scatterplot may suggest similarity among values of the same dataset or correspondence among points of different datasets. Moreover, this correlation may be perceived without drawing any connecting line. This technique is particularly powerful since it could also be used to link together an arbitrary number of scatterplots, both different or complementary, such as in the case of a scatterplot matrix, without cluttering or visibly degrading any of them. This solution can increase significantly the effectiveness of such visualization with respect to the sum of the individual unlinked scatterplots. Colors can also be used to enhance the perception of a variable already displayed by another effect (such as an axis). Figure 6 shows a scatterplot that displays an additional variable through colors.

Size

A further option to provide an additional dimension to the scatterplot is to vary the size of the points. Anyway, this option may lead to occlusion problems if the plot does not provide proper scaling on the two axis. Figure 7 shows a scatterplot with a variable mapped on the size of the points.

Figure 7: a scatterplot with a variable mapped on the size of the points.

Shape

Another approach is to add a third dimension changing the shape of the points. Instead of using only points, each element of the dataset could be drawn as different kinds of glyphs depending on a third variable. This option leads to further possibilities in terms of the paradigm used to choose the shape. One option is to display the points as 'flowers', relating the variable to the number of 'petals' to display. Another option is to display polygons and relating the number of sides to the variable. Moreover, various glyphs, clearly distinct among them, could be used to represent different datasets. Figure 8 shows a scatterplot that uses the shape of the points to display additional information.

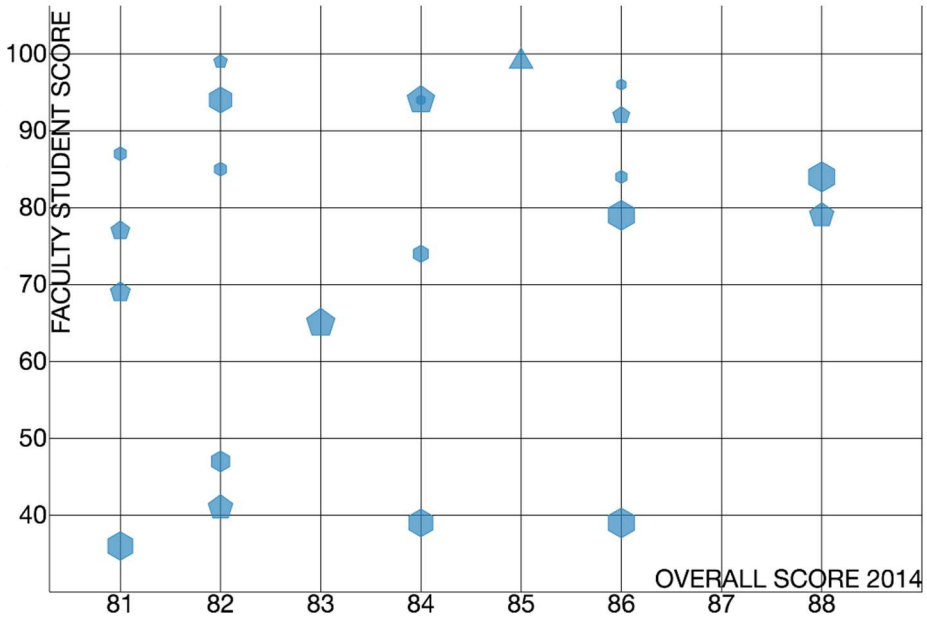

Figure 8: a scatterplot that uses the shape of the points to display additional information.

Orientation

Another possibility when displaying points as shapes is to represent a third dimension changing the orientation of the shape. Usually a dot or line is drawn orthogonally to the perimeter of the shape to better identify the reference point for the orientation. Figure 9 shows a scatterplot that displays an additional variable through the orientation of the points.

Error Bars

The uncertainty is the variability related to a specific variable of the dataset for each point. It provides a generic idea of how precise the measurement of the reported value is, or how far from the recorded value the real value might be. This information is usually reported through error bars if it is related to a variable mapped on the x or y axis (or both). Figure X shows three examples of error bars. Error bars require additional space around the points to be correctly displayed due to the chance of overlapping between points. For this reason they are usually adopted only if the points of the scatterplots are very scattered and occlusion does not occur. Otherwise, due to the space needed to draw the bars, their use would greatly affect the understandability of the representation. As a

result, the use of error bars limits the number of different graphical effects that could be combined on the same scatterplot and should be avoided when displaying more than three or four variables.

Adding more dimensions concurrently

It is possible to use simultaneously more than one of these techniques, independently, to obtain even high visual dimensionality. Figure 10 shows an example of such a scatterplot. However, this is recommended only if the graphical effects are clearly distinguishable, otherwise the visual clarity and benefits of displaying more dimensions at the same time will promptly worsen. Many studies, like the one by (Demiralp et al., 2014), have been carried out to understand how visualization design can benefit from taking into consideration perception, as different assignments of visual encoding variables such as color, shape and size could strongly affect how viewers understand data.

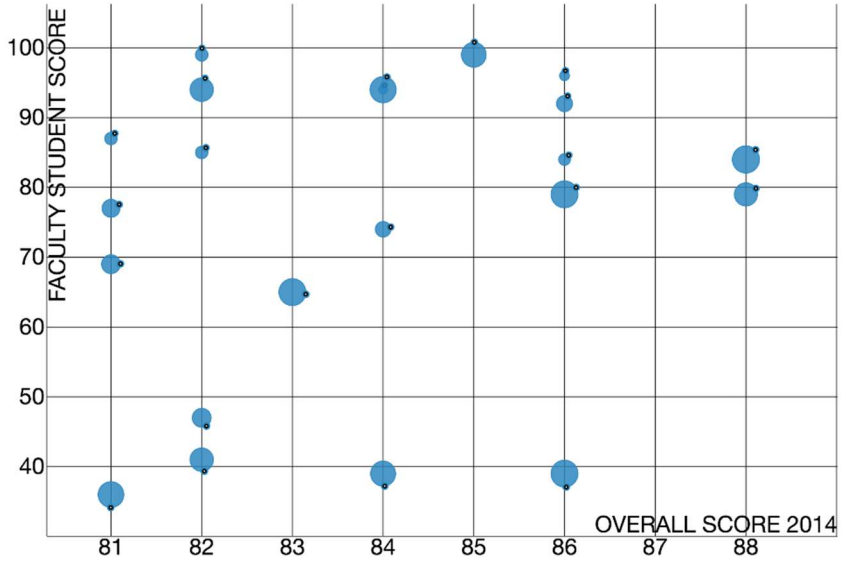

Figure 9: a scatterplot that displays an additional variable through orientation.

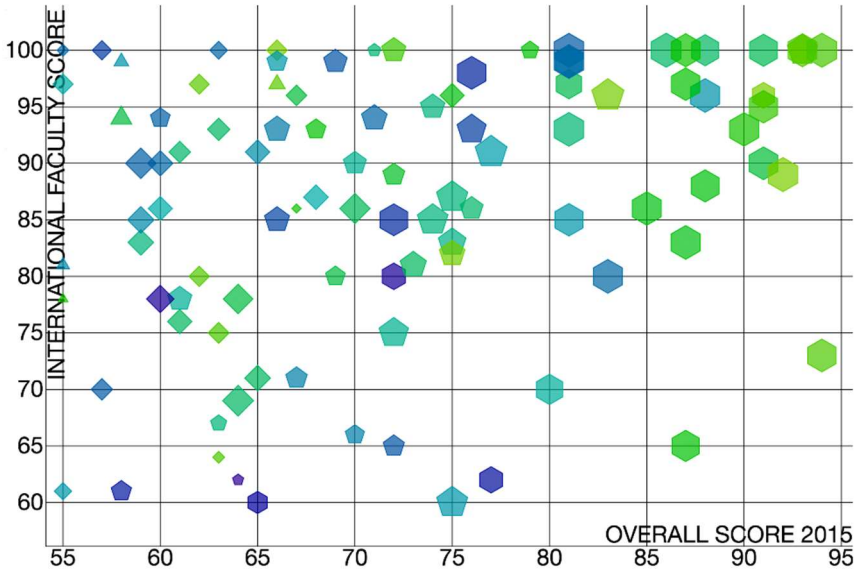

Figure 10: a scatterplot that displays multiple variables through different effects.

Dynamic Visualizations

Even if scatterplots are typically used to display static data, nevertheless they can be very useful when applied to display data that could change dynamically, moreover if the change may be controlled by the user. More complex graphical effects such as animation may be adopted in this case to enhance the comprehension of data as they change over time. This is the case of data characterized by one or more time-related variables, such as stocks values in finance or weather conditions in forecasting.

Scatterplot Matrix

The simplest approach to adapt the scatterplot to multivariate data is to produce a series of scatterplots for each pair of variables and display them together on a single screen or page. This visualization technique is called Scatterplot Matrix and for k variables it requires k(k-1)/2 pairs and therefore scatterplots. Unfortunately, this solution presents a major problem: analyzing all the scatterplots may require a lot of time, depending on the number of variables, thus this solution is not optimal when dealing with time-related tasks. To overcome this problem, different visualization techniques may be adopted to interact with the dataset and simplify data comprehension. Figure 11 shows an example of scatterplot matrix.

Brushing

Brushing is the action of selecting a subset of the points displayed on the scatterplot. Four brushing operations have been defined by (Becker and Cleveland, 1987): highlight, shadow highlight, delete, and label. To perform these operations, it is necessary to resize a rectangle, called the brush, over one of the scatterplots. The corresponding points on each different scatterplot are then affected by the chosen operation. The brush can be moved to different regions of the scatterplot by moving the mouse. At any time, the user can stop the brushing operation, change the shape of the brush or the chosen operation and then resume the brushing.

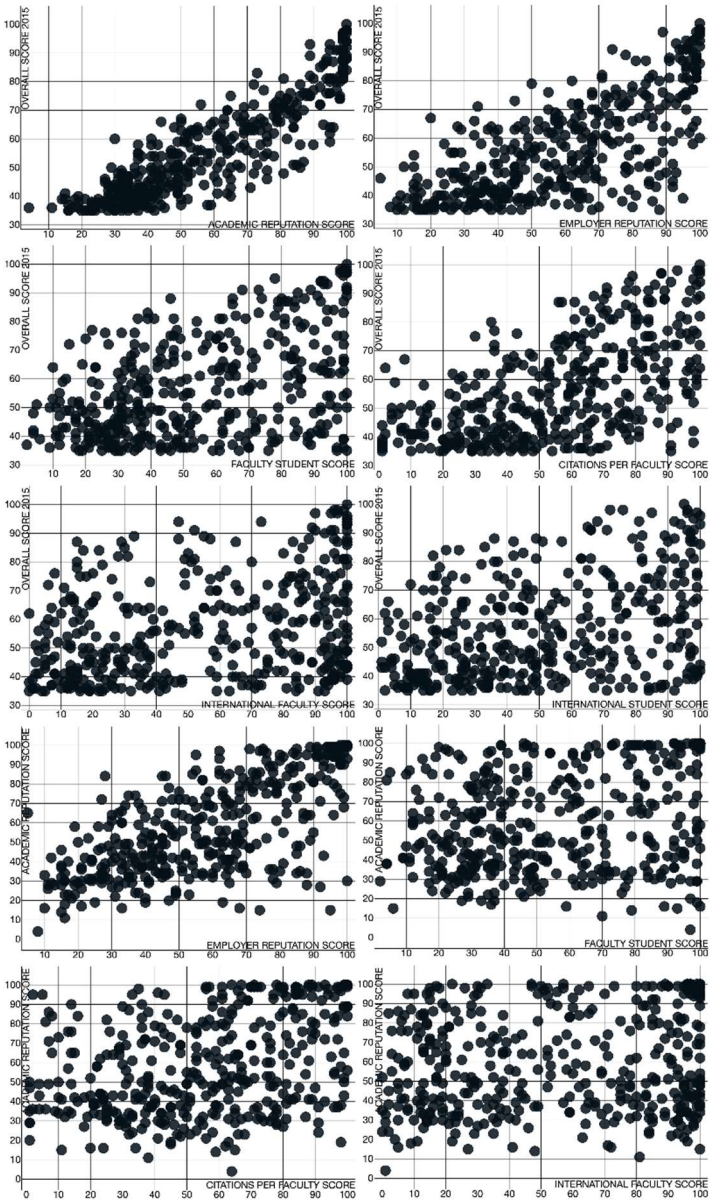

Figure 11: an example of scatterplot matrix.

Dimension Reordering

One of the problems when dealing with scatterplot matrixes is to simplify the understandability of the data. One possibility is to change the way the scatterplots are displayed and ordered to enhance the presence of clusters, patterns or trends. Different approaches have been investigated and adopted, such as the systematic dimension reordering approach of (Ankerst et al., 1998) where similar dimensions in a multidimensional dataset are placed next to each other. Using a scatterplot matrix it is possible to order independently rows and columns. In the systematic dimension reordering approach, similarity are displayed on the column and dissimilarity on the row order.

3D Scatterplots

Another way to display multidimensional data through scatterplots consists in adopting a 3D visualization. 3D scatterplots exploit the third dimension, representing three data dimensions on the x, y and z coordinates, in a three-dimensional space. The third dimension allows the user to interact with the scatterplot to change the viewport (with two or three degrees of freedom). Hypothetically, more coordinates could be added to the model, leading to an n-dimensional spatial representation. Since 3D scatterplots are represented on displays as 2D images, the 3D representation needs to provide useful hints to properly display depth and avoid occlusions or misinterpretations of data. Occlusions can be addressed also in 2D representations by using another data dimension for depth sorting. The latter can also be compared to a full 3D scatterplot where the only difference is the missing rotational interaction in 3D. This mapping also requires three axis: two for spatial positions and one for sorting. 3D scatterplots make it possible to obtain more flexibility in the data mapping simply avoiding to fix certain data dimensions to only certain specific scatterplot axis: this could be obtained allowing the user to exchange the dimensions mapped on each axis, either by swapping the dimension of one or two axis or by manipulation of dimensions. 3D scatterplots may also consist of more complex versions, including additional graphical effects (color, size, orientation, shape, etc.) to represent additional information related to the displayed data, guideways (reference lines from the data up to some reference points) and combinations of scatter data with additional objects as fit surfaces. A common application of the 3D scatterplot is to show both experimental and theoretically-adjusted data in order to be able to determine the points of agreement. In figure 12, a scatterplot can be observed in three dimensions that makes use of the size of the spheres to map an additional attribute. Overall, 3D scatterplots have certain advantages and limitations with respect to 2D models, as depicted by (Escudero et al., 2007).

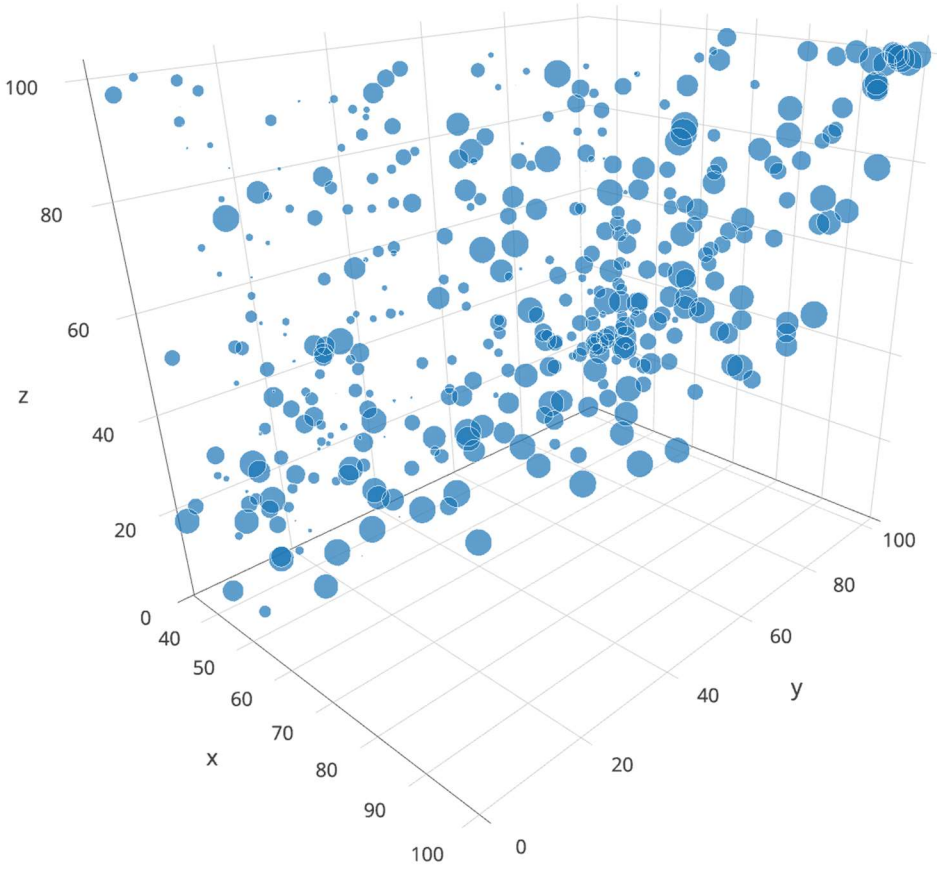

Figure 12: a 3D scatterplot displaying an additional variable through size.

Advantages

In a 3D scatterplot, maintaining the same density of points as in a 2D scatterplot involves increasing the number of experimental data to be displayed (larger sample space). If the number of points of the initial 2D scatterplot is maintained, there is a greater discrimination of the relations between variables, since a characteristic is added to the data. The use of 3D scatterplots with volume visualizations for the glyphs to represent the data provides the possibility of using procedural techniques to generate the forms (Ebert et al., 2000). These techniques allow the user to increase the number of dimensions of the data to be shown by exploiting the shape of the glyphs, thus taking advantage of the pre-attentive ability of the human visual system to discriminate forms. To obtain the best

result from a 3D scatterplot, it is necessary to achieve an efficient attribute mapping and to provide the necessary interaction tools to navigate and examine the data: these requirements enhance the expressive power of a 3D scatterplot and allow the user to analyze complex relationships among multiple variables.

Limitations

It is not advisable to abuse multidimensionality if it is not absolutely necessary and the result is not visually illustrative. Moving information representations from 2-dimension to 3-dimension is not a simple task, since the extra dimension may greatly affect how information can be presented and interpreted. The visualization must make an efficient use of the additional dimension and avoid that the new representation is misinterpreted by the user as a consequence of an inappropriate mapping. Special consideration must be given to the perception of spatial distance. The size of the objects can cause the user to not perceive the correct perspective of the information shown: it is difficult to discriminate among the different depths of the objects and to address this problem it is necessary to provide the appropriate interactions tools. A disadvantage arising from the use of three-dimensional objects is occlusion, which occurs when one object covers another or occupies the same spatial position for two coordinates in the 3D representation. This type of problem occurs mainly when the density of data items to be displayed is large, or when simply a very large object is positioned in front of smaller objects.

Remarks

The reason behind such a various enumeration of scatterplot solutions is that none of them could be considered the best version: each implementation could be less or more useful depending on the specific task the user intends to solve. Eventually, more than one kind of scatterplot should be used for the same dataset to address different tasks. Overall, a simple classification could distinguish among 3D scatterplots, scatterplot matrices and standard scatterplots with additional dimensions. 3D scatterplots are more useful when dealing with a huge amount of data with a dense distribution on the x and y axis, allowing the user a better analysis through spatial navigation. Scatterplot matrices are more useful when the task is to search for correlations between two variables of the dataset: each scatterplot of the matrix may display two variables, and the user just need to analyze them all, one by one. For other tasks, the best solution is adding dimensions to the standard scatterplot, as different graphical effects provide a better insight on the data depending on visual perception criteria, as investigated by (Demiralph et al., 2014) and many others.

References

Miller, J.: Earliest Known Uses of Some of the Words of Mathematics. http://jeff560.tripod.com/mathword.html (1995). Accessed 15 January 2017

Valiati, E.A.R.: Taxonomia de Tarefas para Técnicas de Visualização de Informações Multidimensionais. Porto Alegre, PPGC/UFRGS, 2005 (Technical Report, in portguese). http://www.inf.ufrgs.br/~carla/papers/EValiati.pdf.

Pillat, R. M., Valiati, E. R., & Freitas, C. M: Experimental study on evaluation of multidimensional information visualization techniques. Proceedings of the 2005 Latin American conference on Human-computer interaction, ACM (2005)

Demiralp, Ç., Bernstein, M. S., & Heer, J.: Learning perceptual kernels for visualization design. IEEE transactions on visualization and computer graphics 20.12 (2014): 1933-1942.

Becker, R. A., & Cleveland, W. S.: Brushing scatterplots. Technometrics 29.2 (1987): 127-142.

Ankerst, M., Berchtold, S., & Keim, D. A.: Similarity clustering of dimensions for an enhanced visualization of multidimensional data. In Proceedings of the IEEE Symposium on Information Visualization. pp. 52-62 (1998).

Escudero, M., Ganuza, M. L., Wilberger, D., & Martig, S. R.: Scatter Plot 3D. In IX Workshop de Investigadores en Ciencias de la Computación (2007).

Ebert, D. S., Rohrer, R. M., Shaw, C. D., Panda, P., Kukla, J. M., & Roberts, D. A.: Procedural shape generation for multi-dimensional data visualization. *Computers & Graphics*, 24:375–384 (2000).